图书情报工作
第 63 卷 第 14 期 2019 年 7 月

基于多源数据的专业领域热点探测模型研究*

王晓光[1] 王宏宇[1] 黄菡[2]
[1] 武汉大学信息资源研究中心 武汉 430072 [2] 中南财经政法大学信息与安全工程学院 武汉 430072

摘要：［目的/意义］面向出版业进行专业领域出版时的选题决策问题，对互联网上公开的资讯动态进行多源整合，通过多维度的情报分析探测专业领域内的热点，实现数据驱动的出版选题决策，为出版业的数字化转型与发展奠定坚实基础。［方法/过程］设计一个情报分析模型，面向出版选题决策进行专业领域的热点探测。模型包含热点发现与热度评价两个过程。热点发现过程，通过词频统计和词增长速度算法对专业领域内的热点进行识别；热度评价过程，从内容层面和传播层面两个维度设计并计算一系列指标，对识别到的热点进行热度评价与排序。［结果/结论］以 2018 年 1 月至 4 月的 36 550 条信息、通讯和技术领域多源中文信息为样本进行热点探测实验，实验结果表明，设计的热点探测模型可以有效地探测专业领域内的热点，辅助出版业科学地进行专业领域选题决策。

关键词： 选题决策 热点探测 热点发现 热度计算 热度评价
分类号： G250
DOI：10.13266/j.issn.0252-3116.2019.14.007

"互联网+"与大数据时代的到来，为传统产业带来了新的机遇与挑战。传统出版业需要适应数字化的浪潮，借助信息技术迅速地获取行业数据、掌握消费者与市场的动态，这要求出版业对热点话题变化和大众阅读风向进行全面有效地探测与分析，为消费者提供有价值的动态内容。图书的选题策划，作为出版流程的前端编辑环节，会由专业领域出版商的策划编辑全面分析市场的发行销售数据以及公开的资讯动态来有效地进行选题决策。通过对多源信息的广泛整合与深度分析，可以完成专业领域热点的探测与分析，从而辅助出版业通过数据驱动进行科学的选题决策，为出版业的数字化发展奠定坚实的基础。目前，出版业内已建设的数字化平台多聚焦于电商销售和自媒体运营等终端营销服务，在选题策划等前端编辑环节内，则缺乏相关信息平台提供有效的数据分析与支撑[1]。即使是在出版业广泛使用的"开卷"数据平台上，也仅在出版商、发行商与零售商间对已出版图书的发行与销售等市场数据进行了监控，而没有扩展到互联网上公开的资讯动态[2]。

对专业领域热点的探测，属于情报分析中的热点探测与舆情分析范畴，研究人员对此进行了广泛的研究[3]。在图书情报领域，对某一专业领域内的研究热点进行探测时，通常采用文献计量、共词分析、词语社区发现等方法[4-6]，而针对政府、企业的资讯动态进行监测与分析时，一般通过词频统计、词语重要性排序等步骤，通过舆情分析完成热点的识别与提取[7-8]。在热度计算方面，词语的重要性及词语出现总量、词语增长率等指标常被用于衡量词语的热度，而对词语热度指标的综合性评价，通常基于群体决策法和德尔菲法进行，通过层次分析完成指标的赋权，再应用模糊方法对热度进行准确、客观的综合性评价[9-10]。针对图书出版行业的选题决策过程，也有研究人员基于图书出版行业的实际数据，进行了选题决策分析模型的研究[1]，该研究从作者、图书、出版商、市场和图书馆等多个角度设计了指标评价体系，构建了出版选题决策模型，并利用图书的发行与销售数量等市场数据以及电商网站与社交平台上的用户评分数据，参照中图分类号及电商网站图书分类中的主题类别，对已发行销售

* 本文系国家自然科学基金面上项目"基于大规模开放科学知识图谱的学科新兴趋势探测研究"（项目编号：71874129）和国家社会科学基金重大项目"基于认知计算的学术论文评价理论与方法研究"（项目编号：17ZDA292）研究成果之一。
作者简介：王晓光（ORCID：0000-0003-1284-7164），教授，博士后，博士生导师；王宏宇（ORCID：0000-0002-5063-9166），博士研究生，通讯作者，E-mail：wanghongyu@whu.edu.cn；黄菡（ORCID：0000-0002-1517-9731），硕士研究生。
收稿日期：2018-12-12 修回日期：2019-03-11 本文起止页码：52-61 本文责任编辑：徐健

的图书按照不同的主题进行了热度统计与分析。

针对专业领域出版的图书选题策划，除去已发行销售图书的市场数据与用户评分数据之外，政府、企业及研究机构的资讯动态也是策划编辑进行选题决策时所要参考的有效数据。因此，如何基于互联网公开的多源资讯动态探测专业领域内的热点，并对热点的热度进行多元、客观地评价与排序，为专业领域出版时的选题策划环节提供有效的数据支撑，是出版业数字化发展与转型的过程中亟待解决的问题。笔者设计了一个专业领域热点探测模型，以政府公报与行业新闻、专业机构在官网、微博、微信上发布的动态以及科技文献等多源信息为基础，通过专业领域热点发现和热点主题热度评价两大过程，实现了对专业领域热点的探测。在模型的热点发现过程，通过 TF-IDF 词频统计和词增长速度算法对专业领域内的热点进行了识别[11]。在模型的热度评价过程，则从内容层面和传播层面两个维度设计并计算了一系列参数指标[12]，通过模糊层次分析法对识别到的热点进行了热度评价与排序。笔者设计的专业领域热点探测模型，聚焦于专业领域出版时的选题决策问题，进行了情报分析应用实践。

1 文献综述

在专业领域内，各类研究机构、政府部门和大型企业会公开发布大量资讯动态。基于这些多源资讯动态进行的热点探测工作，是一种综合的情报分析过程。对于研究机构的动态，需要依据专业领域内的科技文献完成学科研究热点的探测。而对于政府、企业等组织的资讯信息，则需要针对这些组织在互联网上公开发布的相关资讯进行舆情热点的识别与分析。因此，笔者从研究热点和舆情热点两个方面对热点探测的相关研究进行了梳理，同时，也对热点热度的计算与评价方法和流程进行了介绍。

1.1 学科研究热点的探测

针对具体学科领域内的研究热点进行探测时，通常以科技文献为研究对象，采用文献计量、词频分析、共词分析等方法进行探测[8]。通过对某学科领域内的科研文献进行词频分析或引文分析等统计计量、对文献集合中的高频关键词或高增长率关键词进行共词分析或聚类分析等关联分析来发现该学科领域内的研究热点[4-6]。常采用文献题录工具 SATI、社会科学统计软件包 SPSS 及引文可视化分析工具 CiteSpace 等软件进行上述分析[13]。虽然学科领域内的研究热点大多会在学术论文中展现出来，但随着各类学科领域相关的数字资源和网络资源的持续性增长，对学科研究热点进行探测的研究对象逐渐扩展到了包含科技文献在内的各类信息资源，对于这些数据源中的领域热点，同样可以利用词频分析、共词分析等方法进行探测[14]。

从学科领域研究内容的层面来划分，领域研究热点可以分为一般流行研究热点和潜在重要研究热点[11]。一般流行热点，往往集中于一些新出现的理论概念和重要技术，研究者数量较多，文献数量相对较大。从总体上，这些文献可以反映一个时期的研究关注热点，该类研究热点可能在其他专业领域中也具有较高的流行度；潜在重要研究热点，往往具有较强的专业性，即便是在同一时期，不同专业领域间的数量差异也很明显。从数量上看，此类研究文献往往并不占有优势。相反，只有那些质量较高的专业研究文献才会对此类热点有较多的关注。所以，被引量较大的文献所具有的关键词应当比被引量较少的文献关键词更能反映潜在重要研究热点。扩展到专业领域内的多源资讯动态信息中，流行热点是指当前已经处于热门状态的热点主题，这些热点在多源科技信息中会占有较多的数量，出现的频次较高，并且政策方面也获得了较大的关注，而潜在热点则是指在最新的科技文献、政府公报、行业新闻与专业机构动态中获得了较大关注度的数据中所具有的一些新的关键词、主题词和概念，在后续有可能转化成为热门主题。

1.2 舆情热点的识别与分析

在进行舆情监测与分析时，一般会从词语、主题、事件等不同的层面完成舆情信息的揭示。在微观层面，通常会将舆情文本中的重要关键词进行排序展示来完成舆情的监测及其后续分析。词语频率统计和词语重要性排序等方法[15]常被用于重要关键词的提取；在中观层面，常采用隐狄利克雷分布(Latent Dirichlet Allocation，LDA)等主题模型[16]与自组织映射(Self-Organizing Map，SOM)等聚类方法[17]，以主题为粒度揭示舆情观点。对舆情文本中的热点进行识别与提取，是舆情监测与分析的基础性工作。

热点识别与提取，通常采用词频统计与重要性排序的方法进行[18]，即从基础数据集中抽取关键词并统计各关键词的出现频率，得到关键词列表后，通过计算词语权重，按重要性从高到低的顺序进行排序，选择一定数量的关键词提取出来。在抽取关键词的过程中，一般需要对原始文本进行中文分词、词性标注等预处理，再依据相关专业领域的主题词表或停用词表选取合适的策略对分词结果进行筛选。

由于某些专业领域内常用的基础词汇，将其作为热点的重要程度不足，所以，为了突显出更为重要的词语，通常选用TF-IDF算法进行词语权重计算[19]。TF-IDF算法能够综合考虑每一个关键词在本数据中出现的次数与在全数据集中出现的频率进行综合词语权重计算，这样可以尽可能消除常见词语对后续热点识别与探测的影响。但对于近期、突发的潜在重要热点而言，由于其关键词的分散度较高，与大部分文本集关键词的差异性较大，因此传统的TF-IDF算法并不十分适用。

针对TF-IDF算法识别弱信号较差的情况，有研究人员设计了单一时间窗口内的主题词增长系数及跨时间窗口间的词汇增长速度等指标来对这类词语的重要性进行量化，取得了一定的效果[7]。在一些研究中，为了更好地揭示舆情监测结果的整体性，还会通过主题模型、语义分析等方法对提取到的关键词进行进一步地聚类[20-21]，通过对各类簇内不同关键词之间的关联挖掘，实现对舆情热点的宏观分析。

1.3 热度计算与评价

在热点热度计算方面，除去词语重要性、词语相关文档数量、词语相关文献被引次数、词语增长率等较为常见的指标可以用于衡量词语热度之外，也有一些其他的指标计算算法来对热度进行量化。例如克林伯格于2002年提出的Burst Detection（突发检测）算法，常用于计算一段时期内相对增长率突然增加的焦点词在文档流中的突发权重指数[22]。克林伯格认为文档的出现并不是平滑增长，而是在一定时间内跳跃式增长的过程。任何文档中的词汇都可以被描述成非活跃状态和bursty状态，并且处于bursty状态的等级根据跳跃的剧烈程度而定。克林伯格基于状态机对在一定时间周期内的文档中的词进行建模，从而产生出词在这段时间内的状态转移序列，即，标示了在不同的时刻下词所处于的状态。其中，非活跃状态对应的状态值是0，其他处于bursty状态下的词则从1开始递增。状态值越大，则说明该词在这段时间内越活跃。因此，Burst指数完全可以用于反映一个时间段内各个热点主题的热度[23-24]。

在完成多项衡量热点热度的指标计算后，需要对这些指标能够反映出来的热点热度进行综合评价，以最终确定热点的排序，而目前在学术界还没有统一的用于热度评价的指标体系。部分学者对网络舆情热度及微博热度的指标体系构建做了大量的研究。在构建评价舆情热度或微博热度的指标体系时，通常从用户特征、信息传播特征以及内容特征3个方面来进行[12]。微博意见领袖的存在，便体现了用户特征对热度的影响。而对于传播特征的热度影响力而言，舆情文本或微博的点赞率、评论率、转发率等传播特性可以最直观地反映出一条文本引起的关注热度。舆情及微博内容本身的特征对热度的影响，体现在其文本内容表现的情感极性、相关话题的文本数量等特征上[25]。在指标确定时，也要同时确定具体指标的量度方案。

指标体系构建完成后，需要对各个指标进行赋权，来对舆情文本或微博的热度进行综合性评价，赋权值的方法通常基于群体决策法并结合德尔菲法的思想，应用层次分析法来确定权值[9]。另外，如果被评价对象的某些评价指标相对模糊，导致无法对评价对象做出明确的结论时，则一般会使用模糊综合评价法来对这些指标进行计算，这种方法以模糊数学为基础，应用模糊关系合成原理，从多因素的角度对评价对象隶属等级进行综合性评价，能够较好地解决评估指标和评估标准模糊的问题，减少人的主观臆断所造成的影响，增强评估结果的准确性和客观性。

2 模型设计

2.1 数据格式

面向出版选题决策的专业领域热点探测模型，需以具体专业领域内大量、最新的多源信息为基础数据。基础数据的采集，对于科技文献信息，一方面要通过专业领域相关的主题词在中国知网、万方数据平台等科技文献服务平台上进行科技论文的检索，将检索到的与专业领域相关的科技论文的标题、摘要及关键词等信息记录到数据库之中，同时还要记录文献的发表时间、作者、机构、引用情况等信息，另一方面，则需要及时录入国家科技项目动态的信息，对项目名称、项目级别、申报书名称、申报人、申报单位、申报书简介和经费额等信息进行采集；对于政府公报与行业新闻信息、专业机构动态信息，则需要专业策划编辑预先指定具体的信息采集来源，如专业领域的政府主管部门、权威行业新闻机构、专业领域内的研究团体和大型企业等机构的官网、微博、微信公众账号等信息发布渠道平台等，再针对这些指定的来源进行政府公报、行业新闻和专业机构动态的标题和内容抓取，此外，对微博和微信公众号文章等信息还要采集其转发量、评论量、点赞量等反映传播广度的数据。模型要求的具体基础数据格式如表1所示：

表 1　专业领域热点探测模型基础数据格式

信息类别与来源		采集规范	
		原始输入数据字段	相关数据字段
科技文献	科技论文	标题、摘要、关键字	时间、作者、作者机构、引用情况
	科技项目	申报书名称、简介	时间、项目名称、项目级别、经费额、申报人、申报机构
政府公报与行业新闻	官方网站	标题、内容	时间、机构名称
	新浪微博	内容	时间、机构名称、转发/评论/点赞量
	微信公众号文章	标题、内容	时间、机构名称、点赞量
专业机构动态	官方网站	标题、内容	时间、机构名称
	新浪微博	内容	时间、机构名称、转发/评论/点赞量
	微信公众号文章	标题、内容	时间、机构名称、点赞量

2.2　热度评价指标

笔者提出的专业领域热点探测模型,利用了层次分析法的思想,通过走访专家调研与访谈,对专业领域热点热度的评价指标进行了多层次的分解,建立了每一层次上的评价论域,最终形成了完善的热度评价指标体系。依据专业领域热点热度的在内容和传播两个评价维度确定了两个评价准则,再通过分析不同评价维度中可获取的具体指标,提出了 8 个具体指标。所有被提出的指标均与专业领域热点的热度呈正相关。具体评价指标体系与指标代表符号如表 2 所示:

表 2　专业领域热点探测模型热度评价指标体系

目标层	准则层	指标层(含指标代表符号)
专业领域热点热度	内容层面(A)	A1 相关文档相对数量－r
		A2 作为特征词的词频占比－f
		A3 突发指数热度－b
		A4 缩放速度－v
		A5 缩放加速度－α
	传播层面(B)	B1 微博/微信发文相对数量－q
		B2 政府公报与行业新闻相对数量－t
		B3 转发/评论/点赞相对数量(相对传播广度)－e

每个指标的具体介绍如下:

(1)内容层面(A)。内容层面,是指对在互联网上可以获取到的与专业领域相关的文献、资讯、新闻、动态等从内容上分析专业领域热点的热度。主要依据在当前周期内采集到基础数据的总数据量、包含相关热点的数据量以及与前几个周期内的数据对比所得到的突发指数、缩放速度、缩放加速度等数据进行评价。

A1 相关文档相对数量:在当前时间周期内采集到的所有基础数据中,与某一候选专业领域热点相关的数据所占的百分比比重。取符号为 r。

A2 作为特征词的词频占比:对于当前周期内采集到的每一条基础数据,都会进行特征词的提取,得到其特征词列表。某一候选热点主题被作为特征词的数据数量除以总数据量的比重成为其作为特征词的词频占比。取符号为 f。

A3 突发指数热度:利用克林伯格提出的突发检测算法,由候选专业领域热点在包含前几个时间周期在内的多个周期内的突发态和非突发态之间的代价收益来计算突发指数。取符号为 b。

A4 缩放速度:某一候选专业领域热点在当前周期内的相关文档数量除以该热点在上一时间周期内的相关文档数量的比。取符号为 v。

A5 缩放加速度:某一候选专业领域热点在当前周期内的相对文档数除以该热点在上一时间周期内的相对文档数的比。取符号为 α。

(2)传播层面(B)。在传播层面,主要依据与专业领域相关的专业机构、政府部门或行业新闻机构在本周期内在新浪微博、微信公众号上的微博、推文的总量和与之相关的转发、评论、点赞的总量,以及政府公报与行业新闻的数量来进行评价。

B1 微博/微信发文相对数量:在当前时间周期内采集到的所有微博、微信公众号文章中,与某一候选专业领域热点相关的数据占全部候选热点微博、微信公众号文章数据总和的百分比比重。取符号为 q。

B2 政府公报与行业新闻相对数量:在当前时间周期内采集到的所有政府公报与行业新闻中,与某一候选专业领域热点相关的数据占全部候选热点政府公报与行业新闻数据总和的百分比比重。取符号为 t。

B3 转发/评论/点赞相对数量(相对传播广度):将在当前时间周期内采集到的所有微博、微信公众号文章中,与某一候选专业领域热点相关的微博、微信公众号文章的转发量、评论量及点赞量的和定义为该热点的传播广度。该热点的传播广度占全部候选热点传播广度总和的百分比比重定义为该热点的相对传播广度。取符号为 e。

2.3 模型结构及探测流程

结合文献综述部分的相关思路和解决方案，最终将专业领域热点探测模型设计分为专业领域热点发现和热点主题热度计算两大过程，专业领域热点探测模型总体结构如图 1 所示：

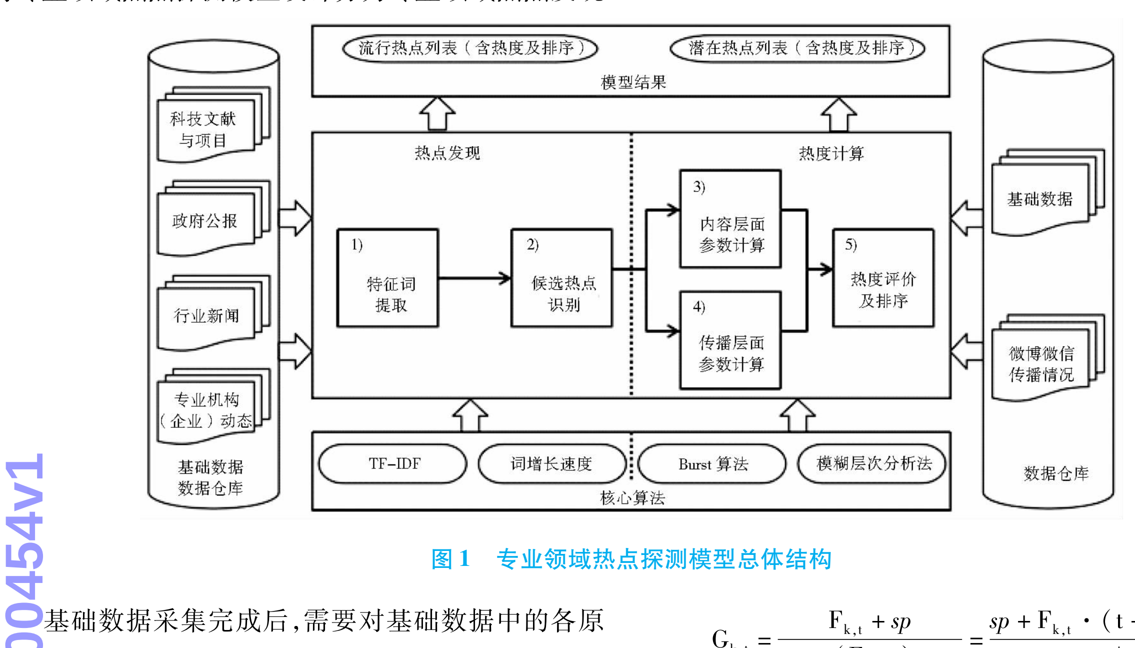


图 1 专业领域热点探测模型总体结构

基础数据采集完成后，需要对基础数据中的各原始输入数据字段进行拼接，将拼接后的字符串经过中文分词、去除停用词后所得的词列表，作为原始输入数据输入到模型中。用 D 来代表当前周期 T = t 内采集到的包含 n 份数据的原始输入数据集合。则 $D = \{D_1, D_2, \cdots, D_n\}$，其中，$D_i$ 表示采集到的第 i 份数据，每份数据中包含若干个词 $\{T_{i1}, T_{i2}, \cdots, T_{ij}, \cdots\}$。

下面，将结合模型的热度评价指标体系以及总体结构，对专业领域热点探测模型的具体流程分步骤进行简要的介绍：

（1）特征词提取。模型的第一步，是对原始输入数据集 D 中的每一条数据 D_i 的词列表提取其特征词列表的过程。首先通过分词、去除停用词等过程对每一条数据的词列表进行过滤，得到每一条数据的特征词列表。之后，再通过 TF-IDF 算法以及词增长速度算法对获得的特征词综合进行词语权重的计算。TF-IDF 算法能够综合考虑每一个特征词在本条数据中出现的次数与在全数据集中出现的频率进行综合的词语权重计算，这样可以尽可能消除常见词语对后续热点识别算法的影响，突显出重要的词语。词增长速度算法可以更加关注于近期突发的潜在重要热点，使这类词语能够被关注到。

依据相关文献[7]，词增长速度算法的公式见公式（1）。其中，$G_{k,t}$ 表示当前周期 T = t 内采集到的原始输入数据集合 D 中某个数据的词列表中的词 k 的词汇增长速度。

$$G_{k,t} = \frac{F_{k,t} + sp}{mean(F_{k,t-}) + sp} = \frac{sp + F_{k,t} \cdot (t - u + 1)}{sp + \sum_u^t F_{k,u}}$$

公式（1）

其中，$F_{k,t}$ 表示时间窗口 T = t 中词汇 k 的词频，t - u + 1 是回溯窗口的大小（即计算的回溯窗口为时间周期 T = u 到时间周期 T = t），$mean(F_{k,t-})$ 表示回溯窗口中词 k 的平均频度，*sp* 是一个平滑系数，由公式（2）给出。

$$sp = \frac{\sum_u^t length(D_u)}{\sum_u^t |V_u|}$$

公式（2）

在公式（2）中，$length(D_u)$ 表示时间窗口 T = u 内采集到的原始输入数据集合的词数量，而 $|V_u|$ 表示时间窗口 T = u 内采集到的原始输入数据集合中包含的词项数（即含有多少个不同的词）。

（2）候选热点识别。模型的第二步，是在特征词列表的集合中进行候选热点的识别，具体步骤如下：将集合 D 中全部数据的特征词列表合并，取构成的新集合中词频最高的若干个词作为候选流行热点。对于潜在热点的识别，需要依据集合 D 中每条数据对应基础数据中的机构名称、作者等相关数据字段对各条数据进行赋权（机构和作者越权威、传播越广泛的数据权重越高，取自然数）。之后，对特征词列表集合中的每条数据重复其权重次数后进行合并，新构成的集合中词频最高的若干个词即为候选潜在热点。

（3）内容层面参数计算。模型的第三步，在得到候选热点主题列表后，要完成对这些候选热点主题的

内容层面参数计算。此处重点介绍模型中用于突发指数热度计算的 Burst 算法，它常用于计算几个周期内相对增长率突然增加的词在文本数据流中的突发指数。对于突发指数，首先利用 Burst 算法对当前周期内候选热点 T_{ij} 的收益 C_{ij} 进行计算，如公式(3)所示[22]：

$$\ln\left[\binom{n}{r'}(\beta \cdot R/N)^{r'}(1-\beta \cdot R/N)^{n-r'}\right] - \ln\left[\binom{n}{r'}(R/N)^{r'}(1-R/N)^{n-r'}\right] \quad 公式(3)$$

其中，β 为算法的经验值，r′表示集合 D 中包含词 T_{ij} 的数据个数，N 表示几个周期内所有数据汇总的集合 D′中的数据个数，R 表示集合 D′中包含词 T_{ij} 的数据个数。突发指数热度通过累加几个周期内候选热点的收益计算而得。

缩放速度是前后两周期内包含候选热点 T_{ij} 的数据个数的比。缩放加速度则是前后两周期内包含候选热点 T_{ij} 的数据个数占该周期内数据总数的比重的比，以消除不同周期内数据总数不同所带来的影响。

(4)传播层面参数计算。模型的第四步，是考虑到在实际情况中，对于专业领域内的热点，除去基于内容层面进行热点热度量化之外，还会考虑到大众传播情况方面的因素。在传播层面，主要计算在当前周期内与候选热点主题词相关的微博、微信公众账号文章、政府公报与行业新闻等几类来源的数量占各类数据中数据全集的比重，由于政府公报与行业新闻这一类别的数据更为严格且更具权威性，所以把政府公报与行业新闻的数量占比分为一类，微博与微信公众号文章的数量归为另一类来计算。另外，对于微博与微信公众号文章这一类数据，从传播广度的角度考虑，还应计算与某一候选热点主题词相关的微博与微信公众号文章的转发、评论、点赞量的相对占比。

(5)热度评价及排序。模型的第五步，需要结合模型第三、四步计算得到的两类层面的热点热度参数指标，通过模糊层次分析法确定具体的评价矩阵，进行两类候选热点热度的综合评价，并完成对候选热点主题列表中热点的排序。由于在内容层面参数、传播层面参数这些热点热度评价指标之中，存在着一些不容易被明确评价的因素，因此，利用模糊语言变量和模糊数可用于量化模糊信息的特点，可以应用层次分析法和模糊综合评价法，从多因素的角度构建评价矩阵，对候选热点主题的热度进行综合性的量化评价。对两类热点主题候选列表分别进行综合评价，得到最终的热点热度。并依据最终的量化评价结果进行排序，便于用户直接对专业领域的热点进行判断，同时，通过对几个周期内不同热点的热度变化进行分析，可以完成专业领域的热点探测和分析。

在出版业机构开展专业领域出版选题决策实践时，为了保证专业领域热点探测模型能够及时、准确地对专业领域内的热点进行探测，模型的多源基础数据及模糊层次分析法确定的评价矩阵等数据需要定期进行更新。

3 实验

3.1 数据准备

笔者选择信息、通讯和技术(Information Communications Technology,ICT)领域，对其 2018 年 3 月和 4 月的选题热点进行了探测实验。由于潜在热点的探测需要专业策划编辑依据基础数据采集时记录的相关数据字段进行大量的权值预设工作，因此，笔者仅就流行热点进行了探测实验。依据模型的基础数据采集规格，利用八爪鱼数据采集器爬取了工信部、科技部等网站上的政府公报数据和新浪、搜狐等门户网站上的新闻数据，并利用微博应用程序接口获取了与“通讯”“科技”等话题相关的微博数据，同时，在知网中以“互联网”“信息技术”等为主题进行检索收集了科技论文数据。此外，为了计算跨时间周期的突发指数热度、缩放速度与缩放加速度等指标，抓取了 2018 年 1 月和 2 月两个时间周期内的相关数据，构成了本文的初始数据集。

为了准确实现 ICT 领域的流行热点探测，对获取到的初始数据集进行了预处理，过滤其中无内容或字数较少的文本，进而形成了包含 36 550 条文本的实验数据集。实验数据的分布情况见表 3。针对该数据集，通过 IKAnalyzer 工具包进行了中文分词，并进一步构建停用词表去除了实验数据中的标点、符号、代词、介词和连词；同时，由于新闻内容中通常包含其来源信息，为了使分词结果更加准确，笔者将“搜狐科技”“新浪科技”等短语一并加入了停用词表。此外，还在特征词提取步骤中，通过构建非特征词表过滤了“公司”“新闻”等无特殊指代的词语，从而保证了最终识别出的候选热点具有更高的可理解性。

表 3 实验数据分布情况 (单位:条)

时间	政府公报及行业新闻	微博微信	科技论文	合计
2018 年 1 月	4 734	2 326	1 525	8 585
2018 年 2 月	2 421	1 458	1 321	5 200
2018 年 3 月	3 393	1 990	1 998	7 381
2018 年 4 月	4 080	6 202	5 102	15 384
合计	14 628	11 175	10 747	36 550

3.2 实验结果

实际对信息、通讯与技术领域进行流行热点探测实验时,要确定热点热度评价指标体系中准则层及各准则内具体指标项间的权重,以便最终得到可量化的热度计算结果。笔者对信息、通讯与技术领域开展专业出版的代表性出版机构——电子工业出版社、人民邮电出版社以及湖北科技出版社的 7 位专业出版从业人员进行了多轮的电话调研与实地访谈,充分了解了专业领域出版的整体业务流程及专业领域选题决策的具体影响因素,通过德尔菲法确定了热度评价指标体系的权值。

由专业出版领域的多位专家对指标的重要性进行多轮比较直到构建出的隶属度矩阵通过一致性检验后,笔者将隶属度矩阵转化为了权值向量并依据各个指标的取值对最终评价目标进行了综合评价。根据信息、通讯与技术领域专家的意见和相关文献[1,12,25],对该专业领域热点热度评价矩阵的权值向量最终确定为:

准则层权值向量:(0.5,0.5)

指标层(A)权值向量:(0.1,0.5,0.3,0.05,0.05)

指标层(B)权值向量:(0.2,0.5,0.3)

取得权重值后,按照笔者提出的模型,对专业领域热点在内容层面和传播层面的参数进行计算,并完成热度计算与排序。按照各个参数指标和最终热度计算结果的大小从高到低排列,得到的 ICT 领域 2018 年 3 月和 4 月前 20 个流行热点及其排序如表 4 和表 5 所示:

表 4 2018 年 3 月 ICT 领域前 20 个流行热点及排序实验结果

排序	相关文档相对数量 r	作为特征词的词频占比 f	突发指数热度 b	缩放(加)速度 v/a	微博/微信发文相对数量 q	政府公报与行业新闻相对数量 t	相对传播广度 e	最终热度计算结果
1	中国	区块链	用户	拍照	智能	企业	视频	OPPO
2	技术	智能	区块链	OPPO	中国	数据	手机	区块链
3	数据	创新	亚马逊	独角兽	行业	技术	用户	拍照
4	服务	技术	汽车	区块链	视频	中国	网络	中国
5	企业	媒体	传播	实效	区块链	服务	人工智能	视频
6	行业	智慧	出版	拍摄	技术	行业	中国	行业
7	智能	苹果	媒体	GIS	创新	人工智能	服务	数据
8	人工智能	金融	算法	施耐德	服务	用户	工业	服务
9	创新	行业	选择	工业园区	企业	区块链	传播	企业
10	区块链	数据	OPPO	金融	人工智能	美国	区块链	技术
11	用户	汽车	视频	实践经验	智慧	美元	教育	拍摄
12	美国	人工智能	拍摄	智慧	数据	创新	学习	人工智能
13	媒体	用户	美国	开发区	手机	智能	数据	金融
14	网络	传统	智慧	算法	传统	智慧	世界	智慧
15	智慧	融资	技术	教育	世界	金融	企业	用户
16	美元	AI	网络	上市	用户	网络	技术	创新
17	传统	阅读	AI	行业	网络	汽车	行业	GIS
18	金融	算法	施耐德	音乐	金融	媒体	金融	施耐德
19	视频	传播	货币	汽车	选择	融资	创新	实效
20	手机	音乐	音乐	中国	媒体	手机	上市	媒体

表 5 2018 年 4 月 ICT 领域前 20 个流行热点及排序实验结果

排序	相关文档相对数量 r	作为特征词的词频占比 f	突发指数热度 b	缩放(加)速度 v/a	微博/微信发文相对数量 q	政府公报与行业新闻相对数量 t	相对传播广度 e	最终热度计算结果
1	科技	区块链	用户	芯片	科技	科技	中国	科技
2	数据	数据	区块链	高校	中国	数据	视频	中国
3	技术	智慧	机器人	图书馆	信息	腾讯	数据	芯片
4	中国	阅读	阅读	链接	智能	中国	科技	高校
5	信息	媒体	汽车	自动化	链接	企业	用户	数据
6	发展	智能	传播	教学	数据	技术	创新	链接
7	服务	技术	模型	模型	智慧	服务	城市	信息

王晓光，王宏宇，黄菡.基于多源数据的专业领域热点探测模型研究[J].图书情报工作,2019,63(14):52－61.

（续表5）

排序	相关文档相对数量 r	作为特征词的词频占比 f	突发指数热度 b	缩放(加)速度 v/a	微博/微信发文相对数量 q	政府公报与行业新闻相对数量 t	相对传播广度 e	最终热度计算结果
8	平台	图书馆	建设	设计	视频	发展	区块链	技术
9	智能	链接	媒体	控制	平台	人工智能	技术	视频
10	企业	科技	管理	头条	创新	用户	发展	自动化
11	系统	用户	算法	阅读	企业	平台	建设	设计
12	创新	网络	视频	信息	技术	美国	数字	发展
13	人工智能	系统	美国	档案	服务	区块链	信息	控制
14	网络	算法	智慧	系统	发展	信息	网络	系统
15	智慧	汽车	技术	算法	建设	美元	金融	智慧
16	用户	人工智能	网络	数据	手机	创新	智能	模型
17	腾讯	创新	信息	智慧	人工智能	智能	平台	平台
18	建设	传播	AI	数字	网络	系统	美国	智能
19	设计	中国	自动化	传播	金融	智慧	服务	建设
20	管理	苹果	发展	视频	管理	网络	手机	创新

通过表4和表5,可以进一步探测分析2018年3－4月间我国ICT领域的流行热点变化趋势,如图2所示：

2018年3月		
排序	热度值	最终结果
1	0.2078	OPPO
2	0.2062	区块链
3	0.1742	拍照
4	0.1661	中国
5	0.1639	视频
6	0.1606	行业
7	0.1591	数据
8	0.1581	服务
9	0.1576	企业
10	0.1542	技术
11	0.1526	拍摄
12	0.1514	人工智能
13	0.1501	金融
14	0.1497	智慧
15	0.1485	用户
16	0.1484	创新
17	0.1476	GIS
18	0.1440	施耐德
19	0.1422	实效
20	0.1415	媒体

2018年4月		
最终结果	热度值	排序
科技	0.2017	1
中国	0.1891	2
芯片	0.1850	3
高校	0.1836	4
数据	0.1799	5
链接	0.1778	6
信息	0.1738	7
技术	0.1629	8
视频	0.1602	9
自动化	0.1596	10
设计	0.1581	11
发展	0.1568	12
控制	0.1537	13
系统	0.1536	14
智慧	0.1531	15
模型	0.1517	16
平台	0.1512	17
智能	0.1508	18
建设	0.1481	19
创新	0.1473	20

上升趋势
下降趋势
4月跌出前20词汇
4月涨至前20词汇

图2　2018年3月至4月ICT领域流行热点变化趋势

通过分析实验结果可以发现,笔者提出的模型较为有效地实现了对流行热点的探测:2018年3月,主打AI拍照的新手机"R15"的发布让手机企业OPPO受到了中国用户广泛的关注,另一方面,在金融行业内快速发展的区块链技术与大数据、人工智能等技术一并被我国的互联网服务企业认定为未来技术发展与创新的重点;2018年4月,科技领域发生的美国对中兴公司的制裁事件,让我国意识到国家需要系统地鼓励高校、科研机构和信息技术行业实现智能芯片的自主设计和自动化建设,让科技创新驱动发展。

4　总结与讨论

笔者提出的面向出版选题决策的专业领域热点探测模型,依据互联网上公开的政府公报与行业新闻、专业机构动态及科技文献等多源资讯动态,结合内容层面、传播层面两方面的要素,通过特征词提取、候选热点识别、内容及传播层面参数计算和热度评价及排序等步骤,完成了专业领域热点的探测与热点的热度计

算，实现了对专业领域热点变化趋势的探测。笔者提出的模型有助于从事专业领域出版的相关人员全面、多维的对专业领域内的热点话题变化和大众传播态势进行自动化地探测与分析，从而实现由以往的以经验主导的选题决策向数据驱动下的科学选题决策的提档升级。同时，本文也为出版流程中的选题策划环节提供了切实可行的数据分析支撑。

笔者聚焦于专业领域出版中的选题决策环节，系统地设计了一个基于公开多源资讯动态构建的专业领域热点探测模型，进行了情报分析应用实践。模型结合了TF-IDF与词增长速度两种算法对候选选题热点进行了词语权重的计算，同时，从两个维度多元客观地设计和选择了突发热度指数、缩放速度及加速度、传播广度等指标参数，构建了面向出版选题决策的热点热度评价指标体系，并最终通过模糊层次分析法，完成了热点热度的评价与排序。通过2018年1月至4月的36 550条信息、通讯和技术领域中文多源科技信息进行的热点探测实验，验证了模型对中文专业领域流行热点的探测效果。

出版机构实际进行专业领域出版时，会经历一个相对复杂的选题决策过程。根据专家调研情况，在策划编辑提交选题建议后，中小型出版社多采用编辑部召开选题讨论会的形式通过群体决策进行选题上报。大型出版社通常会采用管理信息系统对多级参与的选题决策业务流程进行管理，提高选题决策过程的流转效率。之后，各出版机构需要将策划完成的选题上报主管行政部门，待获得行政审批后，结束整体的选题决策流程。笔者提出的模型在后续可以用于中文多源信息的自动化采集、处理与分析，实现选题热点的自动探测，从而减轻专业领域出版从业人员对相关选题资料和信息收集、处理与分析的工作量。另一方面，笔者提出的模型也为开展后续专业领域的产业趋势分析和热点追踪咨询奠定了良好的服务基础。

本文也存在着一些不足。例如，模糊层次分析法是一种依赖于群体决策和专家打分的评价方法，因此，不可避免地具有一些主观性的缺点；同时，本文未对探测到的专业领域热点进行进一步聚类分析和关联分析，专业领域出版从业人员进行实际选题决策时仍需一定的人工介入。今后，本研究将尝试基于机器学习的算法，利用历史的市场数据训练回归模型来科学、自动地确定各评价指标的权重，使热度的量化计算结果更为客观精准；此外，也将通过聚类与关联分析模型对选题热点的探测结果进行深入挖掘，进一步为专业领域的出版选题决策过程提供更为有效的数据支撑。

参考文献：

[1] 曾文，徐红姣，车尧，等. 基于图书出版行业大数据的选题决策分析模型研究[J]. 情报学报，2018，37(8)：813－821.

[2] 黄震. 数字化领航传统出版迈入新时代[J]. 出版广角，2018(17)：35－37.

[3] 王洪伟，高松，陆颋. 基于LDA和SNA的在线新闻热点识别研究[J]. 情报学报，2016，35(10)：1022－1037.

[4] ASATANI K，MORI J，OCHI M，et al. Detecting trends in academic research from a citation network using network representation learning[J]. Plos One，2018，13(5)：e0197260.

[5] SALMERON-MANZANO E，MANZANO-AGUGLIARO F，ENERGIES，et al. The electric bicycle：worldwide research trends[J]. Energies，2018，11(7)：1894.

[6] 程齐凯，王晓光. 一种基于共词网络社区的科研主题演化分析框架[J]. 图书情报工作，2013，57(8)：91－96.

[7] 庄婷婷，王平，程齐凯. 一种时间情境依赖的微博话题抽取方法[J]. 信息资源管理学报，2013(3)：40－46.

[8] 陈武，陆伟，韩曙光. 专家检索及热点探测系统设计与实现[J]. 情报杂志，2009，28(12)：113－117.

[9] WANG H，LIU C，ZHAO Z，et al. Efficiency evaluation of an Internet Plus University Student Affairs System based on fuzzy theory and the analytic hierarchy process[J]. Journal of intelligent & fuzzy systems，2016，31(6)，3121－3130.

[10] 杨春静，程刚. 科技情报机构知识服务能力评价体系研究[J]. 情报理论与实践，2017，40(7)：43－49.

[11] 李树青，白云. 基于时序关键词热点识别方法的图情学科研究趋势分析(2000－2009)[J]. 现代图书情报技术，2011，27(5)：69－76.

[12] 何跃，蔡博驰. 基于因子分析法的微博热度评价模型[J]. 统计与决策，2016(18)：52－54.

[13] 李信，李旭晖，陆伟，等. 大数据驱动下的图书情报学科热点领域挖掘——面向WOS题录数据的实证视角[J]. 图书馆论坛，2017，37(4)：49－57.

[14] 陆伟，彭玉，陈武. 基于SOM的领域热点主题探测[J]. 现代图书情报技术，2011，27(1)：63－68.

[15] 郑魁，疏学明，袁宏永. 网络舆情热点信息自动发现方法[J]. 计算机工程，2010，36(3)：4－6.

[16] 陈晓美，高铖，关心惠. 网络舆情观点提取的LDA主题模型方法[J]. 图书情报工作，2015，59(21)：21－26.

[17] 杨于峰，余伟萍，田盼. 基于SOM神经网络的品牌丑闻微博传播分类预测研究[J]. 情报杂志，2013(10)：23－28.

[18] 吴晓娟. 基于微博文本的网络舆情主题演化分析——以"蓝色钱江放火案"为例[D]. 南京：南京大学，2018.

[19] 刘海峰，于利军，刘守生. 一种基于类别分布信息的文本特征选择模型[J]. 图书情报工作，2013，57(15)：137－141.

[20] YU B，YU Y H. Auto-Tracking controversial topics in social-media-based customer dialog：a case study on starbucks[C]//GOBINDA C，JULIE M. Lecture notes in computer science，volume10766.

Berlin: Springer,2018:87－96.
[21] LI N, WU D. Using text mining and sentiment analysis for online forums hotspot detection and forecast[J]. Decision support systems, 2010,48(2), 354－368.
[22] KLEINBERG J. Bursty and hierarchical structure in streams[J]. Data mining & knowledge discovery, 2003, 7(4):373－397.
[23] 高永兵，杨贵朋，张娣,等. 基于突显词博文聚类的官微事件检测方法[J]. 数据分析与知识发现, 2017, 1(9):57－64.
[24] 杨选辉，蔡志强. 基于突变检测与共词分析的关联数据新兴趋势探测[J]. 情报科学, 2018, 36(11): 164－168.
[25] 孙飞显，程世辉，靳晓婷,等. 政府负面网络舆情热度定量评价方法——以新浪微博为例[J]. 情报杂志, 2015(8):137－141.

作者贡献说明：

王晓光:提出研究思路、论文修改与润色；
王宏宇:设计研究方案、设计并进行实验、论文起草及修改；
黄菡:采集实验数据,进行实验。

Towards Professional Publishing: Research on Hotspot Detection Model Based on Multi-source Data

Wang Xiaoguang[1]　Wang Hongyu[1]　Huang Han[2]
[1] Center for Studies of Information Resources, Wuhan University, Wuhan 430072
[2] School of Information and Safety Engineering, Zhongnan University of Economic and Law, Wuhan 430072

Abstract: [**Purpose/significance**] In order to solve the problem of topic selection for professional fields in publishing industry, this paper integrates multisource dynamic information on the Internet to detect the hotspots for professional fields through multi-dimensional intelligence analysis. The data-driven topic selection is realized to lay a solid foundation for the digitization transformation and development of publishing industry. [**Method/process**] A intelligence analysis model towards topic selection was proposed to detect hotspots in professional fields. The model was divided into two steps: the hotspot discovery and the hotness evaluation. The hotspot discovery in this model identified hotspots in professional fields through word frequency statistics and the algorithm of word growth rate. Then, in the step of hotness evaluation, a series of indices in the dimension of content and spread were designed to calculate and evaluate the hotness of the hotspots identified in the last step. [**Result/conclusion**] A hotspots detecting experiment was conducted with 36,550 pieces of Chinese multisource dynamic information in the area of ICT collected from January to April of 2018, which verified the effectiveness of the proposed model. This model can be used in publishing industry to complete the step of topic selection scientifically.

Keywords: topic selection　hotspot tracking　hotspot detection　hotness calculate　hotness evaluation